CSC 5930/9010: Text Mining

Dr. Paula Matuszek
Paula.Matuszek@villanova.edu
Paula.Matuszek@gmail.com
(610) 647-9789
Red Tape

- Web Page: www.csc.villanova.edu/~matuszek
- Syllabus
- Grading
- Academic honesty
- Student Questionnaire
- Today: general overview of the kinds of things we will be covering
The Flood of Information

- Enormous amount of information on the web
  - In 1998, Google indexed 26 million pages.
  - In 2000, Google reached a billion pages.
  - In 2008, Google hit 1 trillion unique URLs.¹

- Equally enormous amount of information in electronic form not on the internet
  - Intranets are moving in the same direction
  - The last physical filing of a drug application to the FDA by one drug company required two trucks to deliver; now applications are filed electronically.

- Result: simple access to text is no longer an issue

¹. http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
Finding Out About

- People interact with all this information because they want to KNOW something; there is a question they are trying to answer or a piece of information they want.

- Simplest approach:
  - Knowledge is organized into chunks (pages)
  - Goal is to return appropriate chunks

- More complex approaches
  - Information requires more than a single chunk
    - Organized pages, visual displays
  - Specific piece of information is retrieved directly
Challenges and Possibilities

• So: Information overload. There’s too much. We would like
  • Better retrieval (especially ranking)
  • Help with handling the documents we have
  • Help finding specific pieces of information without having to read each document

• What might help?
  • Statistical techniques
  • Natural language processing techniques
  • Knowledge domain based techniques
Text Mining

• Common theme: information exists, but in unstructured text.

• Text mining is the general term for a set of techniques for analyzing unstructured text in order to process it better or facilitate "finding out".
  • Document-based
  • Content-based
Document-Based

- Techniques which are concerned with documents as a whole, rather than details of the contents
  - Document retrieval: find documents
  - Document classification: sort documents into known groups
  - Document clustering: group documents into similar classes which are not predefined
  - Visualization: visually display relationships among documents
Document Retrieval

• Document retrieval
  • Return appropriate documents, documents “like this one”

• Examples:
  • Search refinement. Search tools
  • Competitive Intelligence: Retrieve documents from the web relevant to “my product”; clipping service.
  • Help Desk: Retrieve problem descriptions which match the current description

• Issues
  • Synonyms/preferred terms
  • Multiple meanings for terms
  • Performance: does it need to be real-time?
Search Engines

• Goal of search engine is to return appropriate chunks

• Steps involve include
  • asking a question
  • finding answers
  • evaluating answers
  • presenting answers

• Value of a search engine depends on how well it does on all of these.
Asking a question

- Reflect some information need
- Query Syntax needs to allow information need to be expressed
  - Keywords
  - Combining terms
    - Simple: “required”, NOT (+ and -)
    - Boolean expressions with and/or/not and nested parentheses
    - Variations: strings, NEAR, capitalization.
  - Simplest syntax that works
  - Typically more acceptable if predictable
- Another set of problems when information isn’t text: graphics, music
Finding the Information

- Goal is to retrieve relevant chunks. Too time-consuming to do in real-time, so search engines index pages.
- Deciding what to index on (e.g., what is a keyword) is a significant issue.
- Search engine basics:
  - Spider or crawler starts at a web page, identifies all links on it, and follows them to new web pages.
  - Parser processes each web page and extracts individual words.
  - Indexer updates hash table which links words to pages
  - Searcher uses hash table to retrieve pages based on words
  - A ranking system decides the order in which to present the documents: their relevance
Balancing Search Factors

- **Recall and Precision**
  - If there are 100 relevant documents and you find 40, recall is 40%.
  - If you find 100 documents, and 10 are on topic, precision is 10%.
  - Which is more important varies with query and with coverage

- **Speed, completeness, timeliness**
  - How fast can you locate and index documents? Answer a query?
  - What percent of the web/available documents do you cover?
  - How many dead links do you have? How long before information is found by your search engine?

- **Ease of use vs power of queries**
  - Full Boolean queries very rich, very confusing
  -Simplest is “and”ing together keywords; fast, straightforward
  - Field-specific searches
Search Engines Now

• We take for granted most of the above.

• More advanced issues:
  • Cache of document, format filters (pdf, postscript, etc), duplicate identification and removal
  • Summary of document.
  • Focused crawling
  • “More like this”
  • Semantic search
    • Text Runner http://openie.cs.washington.edu/
    • Cognition http://www.cognition.com/
Handling Documents

- Okay, your search engine has returned 50,000 hits. Now what?

- Text mining techniques at document level.
  - Cluster documents to see trends and overall "shape" of the knowledge
  - Classify documents into groups you already know

- Text mining techniques at content level
  - Summarize documents
  - Extract specific information from documents
Document Classification

• Document classification
  • Assign documents to pre-defined categories

• Examples
  • Process email into work, personal, junk
  • Process documents from a newsgroup into “interesting”, “not interesting”, “spam and flames”
  • Process transcripts of bugged phone calls into “relevant” and “irrelevant”

• Issues
  • Real-time?
  • How many categories/document? Flat or hierarchical?
  • Categories defined automatically or by hand?
Document Classification

- Usually
  - relatively few categories
  - well defined; a person could do task easily
  - Categories don't change quickly

- Flat vs Hierarchy
  - Simple classification is into mutually-exclusive document collections
  - Richer classification is into hierarchy with multiple inheritance
    - broader and narrower categories
    - documents can go more than one place
    - Merges into search-engine with category browsers
Classification -- Automatic

- Statistical approaches similar to search engine
- Set of “training” documents define categories
  - Underlying representation of document is bag of words/TF*IDF variant
  - Category description is created using neural nets, regression trees, other Machine Learning techniques
  - Individual documents categorized by net, inferred rules, etc
- Requires relatively little effort to create categories
- Accuracy is heavily dependent on "good" training examples
- Typically limited to flat, mutually exclusive categories
Classification: Manual

- Natural Language/linguistic techniques
- Categories are defined by people
  - underlying representation of document is stream of tokens
  - category description contains
    - ontology of terms and relations
    - pattern-matching rules
  - individual documents categorized by pattern-matching
- Defining categories can be very time-consuming
- Typically takes some experimentation to "get it right"
- Can handle much more complex structures
- Not common currently.
- Note that classification itself is still automatic. Building a directory by hand isn't text mining.
Document Clustering

- Document clustering
  - Group documents based on similarity

- Examples
  - Group samples of writing in an attempt to determine author(s)
  - Look for “hot spots” in customer feedback
  - Find new trends in a document collection (outliers, hard to classify)

- Getting into areas where we don’t know ahead of time what we will have; true “mining”
Document Clustering -- How

Typical process is:

- Describe each document
- Assess similarities among documents
- Establish clustering scheme which creates optimal "separation"

One typical approach:

- Document is represented as term vector
- Cosine similarity for measuring association
- Bottom-up pairwise combining of documents to get clusters

Assumes you have the corpus in hand
Document Clustering

- Approaches vary a great deal in
  - document characteristics used to describe document (linguistic or semantic? bow?)
  - methods used to define "similar"
  - methods used to create clusters

- Other relevant factors
  - Number of clusters to extract is variable
  - Often combined with visualization tools based on similarity and/or clusters
  - Sometimes important that approach be incremental

- Useful approach when you don't have a handle on the domain or it's changing
Document Visualization

- Visualization
  - Visually display relationships among documents

- Examples
  - hyperbolic viewer based on document similarity; browse a field of scientific documents
  - “map” based techniques showing peaks, valleys, outliers
  - graphs showing relationships between companies and research areas

- Highly interactive, intended to aid a human in finding interrelationships and new knowledge in the document set.
Content-Based Text Mining

• Methods which focus in a specific document rather than a corpus of documents
  • Document Summarization: summarize document
  • Feature Extraction: find specific features
  • Information Extraction: find detailed information
  • Sentiment Analysis: find the mood or tone of a document
• Often not interested in document itself
Document Summarization

- Document Summarization
  - Provide meaningful summary for each document
- Examples:
  - Search tool returns “context”
  - Monthly progress reports from multiple projects
  - Summaries of news articles on the human genome
- Often part of a document retrieval system, to enable user to judge documents better
- Surprisingly hard to make sophisticated
Document Summarization -- How

- Two general approaches:
  - Extract representative sentences/ clauses
  - Capture in abstract representation, generate summary
- Representative sentences/ clauses
  - If in response to search, keywords. Easy, effective
  - Otherwise TF*IDF, position, etc;
  - Broadly applicable, gets "general feel"
- Capture and generate
  - Create "template" or "frame"
  - NL processing to fill in frame
  - Generation based on template
  - Good if well-defined domain, clearcut information needs
Feature Extraction

- Group individual terms into more complex entities (which then become tokens)

- Examples
  - Dates, times, names, places
  - URLs, HREFs and IMG tags
  - Relationships like “X is president of Y”

- Can involve quite high-level features: language

- Enables more sophisticated queries
  - Show me all the people mentioned in the news today
  - Show me every mention of “New York”

- Also refers to extracting aspects of document which somehow characterize it: length, vocab, etc
Feature Extraction -- How

- Human-meaningful features: Parse token stream, applying pattern-matching rules
  - general, broadly applicable features (dates)
  - domain-specific features (chemical names)
  - Can involve very sophisticated domain knowledge.

- Statistical features:
  - document length, vocabulary used, sentence length, document complexity, etc, etc

- Often first step in document-based analysis such as classification
Information Extraction

• Retrieve some specific information which is located somewhere in this set of documents.

• Don’t want the document itself, just the info.
  • Information may occur multiple times in many documents, but we just need to find it once
  • Often what is really wanted from a web search.

• Tools not typically designed to be interactive; not fast enough for interactive processing of a large number of documents

• Often first step in creating a more structured representation of the information
Some Examples of Information Extraction

• Financial Information
  • Who is the CEO/CTO of a company?
  • What were the dividend payments for stocks I’m interested in for the last five years?

• Biological Information
  • Are there known inhibitors of enzymes in a pathway?
  • Are there chromosomally located point mutations that result in a described phenotype?

• Other typical questions
  • who is familiar with or working on a domain?
  • what patent information is available?
Information Extraction -- Semantic

- Create a model of information to be extracted
- Create knowledge base of rules for extraction
  - concepts
  - relations among concepts
- Find information
  - Word-matching: template. "Open door".
  - Shallow parsing: simple syntax. "Open door with key"
  - Deep Parsing: produce parse tree from document
- Process information (into database, for instance)
- Involves some level of domain modeling and natural language processing
Sentiment Analysis

- Determine the overall tone or mood of a document: positive, neutral, negative.

- Actual content of the document is not the focus; may provide a summary from multiple documents.

- Documents are often short, social media such as Twitter

- More advanced may detect emotions such as angry, sad, happy.

- Currently popular for
  - analyzing social trends
  - looking at political events
  - market analysis
Sentiment Analysis --How

• Semantic
  • Lists of polar words: good, bad, etc
  • Get a count or average for the document
  • Needs some sophistication to deal with negation and sarcasm

• Statistical/Machine learning
  • A standard classification problem
  • Provide multiple instances of documents already scored as positive, neutral, negative and let the system figure it out.
Why Text Is Hard

- Natural language processing is AI-Complete.
- Abstract concepts are difficult to represent
- LOTS of possible relationships among concepts
- Many ways to represent similar concepts
- Tens or hundreds or thousands of features/dimensions

http://www.sims.berkeley.edu/~hearst/talks/dm-talk/
I saw Pathfinder on Mars with a telescope.

Pathfinder photographed Mars.

The Pathfinder photograph mars our perception of a lifeless planet.

The Pathfinder photograph from Ford has arrived.

When a Pathfinder fords a river it sometimes mars its paint job.¹

¹ http://people.ischool.berkeley.edu/~hearst/talks/dm-talk/sld009.htm
Why Text is Easy

• Highly redundant when you have a lot of it
• Many relatively crude methods provide fairly good results:
  • Pull out “important” phrases
  • Find “meaningfully” related words
  • Create summary from document
  • “grep”
• Evaluating results is not easy; need to know the question!
Summary: This Course

- Broad field, moving fast
- Draws on aspects of
  - machine learning techniques
  - natural language processing
  - knowledge representation and ontologies
- Course will cover intro to major topics
- Labs will include some hands-on experience with several tools, especially NLTK and GATE.
- Presentations and projects will provide more depth
- If you’re interested in a specific topic -- do a project!